BIG DATA IN THE DIGITAL HUMANITIES. NEW CONVERSATIONS IN THE GLOBAL ACADEMIC CONTEXT

Antonio Rojas Castro @RojasCastroA

Antonio Rojas Castro earned a doctorate in the humanities from the Universitat Pompeu Fabra (2015, Barcelona). Also at this university he was a pre-doctoral fellow, an FPI grantee belonging to the Todo Góngora II research group and a lecturer on academic writing and literary studies subjects. In 2015 he was joint editor of a monograph on the Digital Humanities for the magazine *Ínsula*. He is currently editor of *The Programming Historian en español*, is in charge of communicatiosn at the European Association for Digital Humanities (EADH) and works as a post-doctoral fellow at the Cologne Center for eHumanities (Germany)

Introduction

Christmas 2016. A perfect time to think back, sum up and publish lists of the main events of the year. Google Trends published the most popular searches grouped into categories such as "News", "People", "Technology", "Films", "Music", "Sport" and "Deaths". A few days earlier the Swedish company Spotify, which provides online access to millions of songs, launched an advertising campaign based on data produced by users. Some of the huge posters plastered all over the streets of London display messages such as: "Dear person who played 'Sorry' 42 times on Valentine's Day, what did you do?"; or "Dear 3,749 people who streamed 'It's the End of the World as We Know It' the day of the Brexit vote, hang in there."

Spotify's campaign is both surprising and effective because it plays on the viewer's engagement. But what has all this got to do with the humanistic disciplines that study documents, texts and images of the past? Or, in other words, how can handling the large amount of data amassed by companies help us gain a better understanding of the limits of our thought, language and historical events – basically all the *expressions of our human mind*?

If we accept that humanistic disciplines such as philosophy, philology and history are characterised not only by a specific object of study but also by a method that seeks to understand particular, unusual and even unique cases through text commentary, then the answer will no doubt be negative: "nothing, or very little". However, as Professor Rens Bod (2013) recently argued, since antiquity humanists have also sought general principles, laws and patterns to explain our culture, and have often (for good or for bad) changed how we perceive the world.

We should begin by dismissing certain clichés about the humanities and ask ourselves about their classic objects of study, bearing in mind the methods that are currently available. This requirement is not unrelated to the work of humanists, who have always been in contact with other fringe disciplines such as anthropology, Marxism and gender studies. Indeed, in recent years humanists have established a fruitful dialogue with computer studies and the social sciences – which has been called a "computational turn" (Berry, 2011). In this academic context, the expression "Big Data" has directly found its way into debates on "scale" – how can we study all the eighteenth- and nineteenth-century novels written in England, France, Germany, the United States or Japan?; or, more commonly, in a cross-cutting way through concepts more familiar to humanists, such as "distant reading" (Moretti, 2007) or "macroanalysis" (Jockers, 2013).

Humanistic disciplines such as philosophy, philology and history are characterised not only by a specific object of study but also by a method that seeks to understand particular, unusual and even unique cases through text commentary.

These changes have been made possible by the fact that statistical and computing methods, as well as other methods related to the social sciences, have been modified and have succeeded in adapting their conceptual models to the complexity of texts (English and Underwood, 2016). In other words, we are dealing with a genuine conversation in which the various interlocutors talk and listen to each other.

Concerning the particular in the universal

The expression "Big Data" has been spreading in the experimental sciences and the media since 2011, as if an increased amount of available data were the next scientific breakthrough. The term is used in academia, industry and the media... but what exactly does it mean? Is it an object of study, a method, a group of technologies or a discipline?

One of the few articles to have shed some light on the matter is entitled "Undefined by Data: A Survey of Big Data Definitions". The authors (Ward and Barker, 2013) collate the various definitions of "Big Data" provided by major technology companies like Oracle, Intel and Microsoft and a few previous reports. In general, the definitions combine two important ideas: storage of a large volume of data (some authors speak of 500 Terabytes per week); and analysing this data quantitatively and visually to find patterns, establish laws and predict conduct.

The classic definition of "Big Data" is a formula that is easy to understand and memorise – the three *Vs*: *Volume* (Terabytes, Petabytes, Exabytes), *Velocity* (data that is constantly generated) and *Variety* (texts, images, sounds) (Ward and Barker, 2013). Some reports have subsequently added a fourth *V*, which stands for the term *Veracity*. Though this volume-based definition of Big Data only makes sense if we consider blogs, the social media and sensors to be main sources of data.

In contrast, the classic object of study of the humanities is usually texts and analogue images which have fortunately been digitised and published in computer-legible format. In other words, if we take the three *Vs* as a basis, we have to admit that we cannot speak of Big Data in the strict sense in the humanities. For one thing, the classic works of Spanish Golden Age poetry fit into a 4GB pen drive; for another, archives and libraries do not constantly produce new data and at a high speed on our poets, writers or artists (or rather, this data is not accessible to researchers). As for variety, we are dealing with image files in TIFF, JPEG or another similar format, and semi-structured text in XML format or, without markup, in TXT format.

Before the advent of Google Books in 2004, digital humanists worked to digitise corpuses of texts and images in the form of digital editions, libraries and files. The European Association for Digital Humanities (EADH) provides a good example of the type of projects carried out. Since 2015, the association has devoted a space on its website to documenting and promoting access to European Digital Humanities projects conducted in the past five years. The initiative is participatory in nature because any researcher (whether or not they belong to the association) can fill in the form available on the website and submit a description of their project providing details of the name of the project, a descriptive summary, collaborating institutions or the team in charge, among other fields. So far, at the time of writing this article, the association has received 175 submissions. If the titles and summaries are analysed with Voyant, a tool for counting the most frequently used words, it is easy to see that the projects abound in words related to the subject of this article, such as "data", "information" and "database", and others that denote the scale or size of the project, including "archive", "collection", "platform" and "library".

Words used most frequently to describe Digital Humanities projects in Europe CC-BY

The current state of the Digital Humanities in Europe can be gauged by three aspects: projects, tools and research groups. Prominent among the projects for making digital texts available online are Oxford Text Archive, Deutsche TextArchive, Eighteenth-Century Poetry Archive, and DigiLibt. Tools for textual analysis include Alcide, CATMA and Stylo R. Infrastructure and research groups such as CLARIN, CLiGS and Electronic Text Reuse Acquisition Project are also important. These initiatives use algorithms to attribute authorship of texts (Burrows, 2002), discover latent themes underlying a large group of texts (Blei, 2012), or detect cases of intertextuality in several authors' literary output (Ganascia, Glaudes and Del Lungo, 2015). Suffice it to say

that many of these procedures are comparable to automatic image processing (Rosado, 2015).

The ultimate aim is usually to find patterns that help understand literary and artistic creations. But text commentary – close reading – continues to play an important role even when statistical methods are used to analyse texts, because researchers shift their attention from the whole to the detail and from the detail to the whole to check that their ideas about the work are correct and accordingly gain a better understanding of the different layers of meaning, the central themes, the events and the style. Put another way, distant reading and close reading are not mutually exclusive because researchers usually combine both strategies: they first gain an overview and then filter and examine the details for a deep comprehension. They usually complete their analysis with visualisations of information in the form of marginal annotations, parallel texts that are connected in some way (colours, density, contrast between form and substance, arrows) or more abstract structures like maps, trees and graphs (Jänicke, Franzini, Cheema and Scheuermann, 2015).

In the humanities we can only speak of Big Data in connection with the technologies associated with this phenomenon, such as data mining, stylometry or natural language processing.

To sum up, although the volume of data is not comparable to that currently generated by the social media, blogs and major companies, in the humanities (and specifically in literary studies) we can only speak of Big Data in connection with the technologies associated with this phenomenon, such as data mining, stylometry or natural language processing.

Data as a human construction

The conversation between the humanities and Big Data does not merely boil down to adopting algorithms for studying large holdings of texts and images quantitatively. Indeed, digital humanists have played an active part in the debates on the nature of data.

In a context in which data is equated with objective, irrefutable evidence, it is constantly stated that data is in fact a human construction; that is, it is conditioned by the time, place, language and ideology of the actors involved in gathering it. For example, the researcher Johanna Drucker (2011) rejects the term "data" – Latin for "that which is given to us" – and uses instead the term "capta" meaning "that which has been taken or collected"; evidently this critical intervention highlights the impartial and incomplete nature of data.

Digital humanists have also stressed the temporality of data – for all data has a date of creation and expiry – and the fallacy of separating data from metadata (that is, data such as title, maker, theme, description, date, format, identifier, source, language, etc.). Actually there is no such thing as second-grade data, as embodied by the root *meta*; metadata is just as important, selective and impartial as data because it is produced by humans (or rather by algorithms designed by human beings). Equally invalid is the distinction – which dates back to Lévi-Strauss's culinary triangle – between "raw data" and "cooked data" or between "data", "raw material" and "information".

Indeed, for researchers like Tom Boellstorff (2013), data is dense, interpretative and contextual, and it is therefore preferable to speak of "thick data". Paraphrasing the anthropologist Clifford Geertz, data should be regarded as "our own constructions of other people's constructions" of objects imagined by a particular community.

For example, the Text Encoding Initiative is a non-profit organisation that publishes *Recommendations* on how to encode humanistic texts with XML markup language so that they

are interchangeable and, more or less, standard. It is a participatory organisation in which any researcher can suggest changes or improvements based on their experience to the set of labels defined by the consortium. Up until 2012, however, none of its members had questioned the fact that the label <*sex*> for describing the sex of a person mentioned in a text complied with standard ISO/IEC 5218:2004 and that the attributes (*@value*) were given as single-digit codes 1 (male), 2 (female), 9 (not applicable) and 0 (not known).

The situation was re-examined when a female researcher pointed out that this typology was sexist, as it put women in second place with respect to men, and codified patriarchal structures with markup language (Terras, 2013). With this I do not wish to detract from the importance of the TEI, especially in giving shape to the Digital Humanities, but rather to stress that technology, data, algorithms and standards are the product of an interpretation of the world and bear cultural marks. In conclusion, data should not be viewed as absolute truths but be questioned critically.

Our cultural heritage is not fully digitised, despite the collective efforts of initiatives like Europeana. Only 23% of European collections have currently been digitised.

In defence of theory

In literature on Big Data it is also common to find that theory is discredited. The argument is basically as follows: if we have large amounts of data and effective statistical methods, we do not need theories, models and hypotheses, which need to be proven or refuted with experiments. Put another way, in the era of the Petabyte, scientific method is obsolete (Anderson, 2008). The dismissal of theories and models has not only been given credit in the business world, but it has also been accepted in a few humanistic writings. Jean-Gabriel Ganascia (2015: 632–33), for example, claims that a theory or previous hypothesis is no longer necessary if we analyse all the existing data as opposed to a sample or small group, as has been done so far.

In contrast to this viewpoint, a considerable number of writings have confirmed the importance of theories, models and hypotheses for research. It should be remembered that our cultural heritage (documents, texts, paintings, images, sounds) is not fully digitised, despite the collective efforts of initiatives like Europeana. According to the latest report issued by the European Commission project ENUMERATE (Nauta and Wietske, 2015), only 23% of European collections have currently been digitised. The survey was answered by some 1,000 European institutions including libraries, museums and archives. These institutions have yet to digitise some 50% of their collections and admit that about 27% of their holdings will not be digitised. These figures highlight the fact that much of our heritage is not accessible on the internet.

Digitisation always involves making a selection based on the resources available to the institution or working group in charge of digitising the documents; but this selection furthermore stems from ideological and identity reasons. It should not be forgotten that museums, libraries and archives are publicly funded institutions and their role is to preserve and disseminate the cultural heritage of a community (for example, a nation). In addition, formats, markup languages and algorithms are also part of a particular culture and ideology and go hand in hand with many assumptions that vary depending on the context.

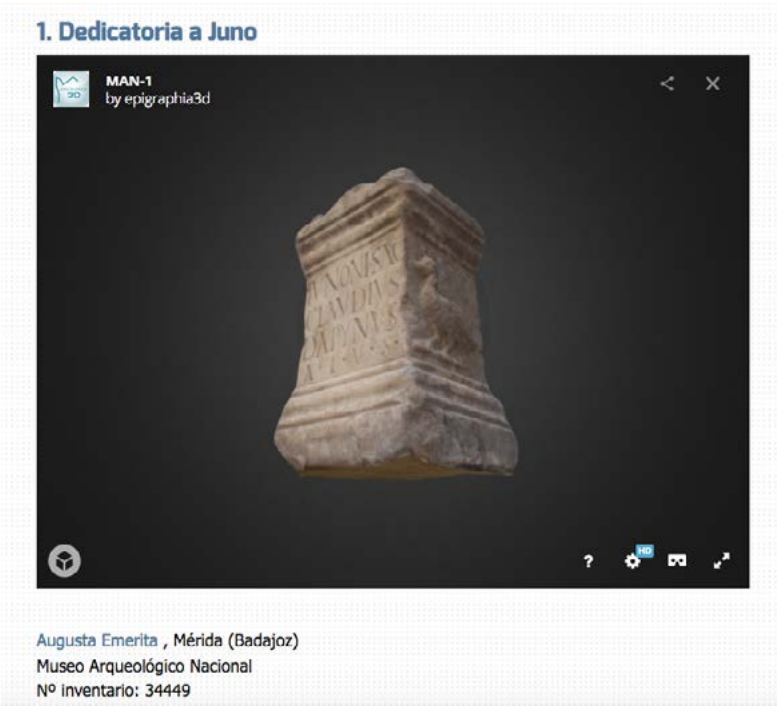
From a humanistic viewpoint, it is thus hard to believe that analysing large amounts of data could renders scientific method useless, because we never have all the existing data – one of the vectors of Big Data is the Velocity with which new data is generated – because the data is

erroneous or ambiguous, or because data processing (automatic or otherwise) is determined by our culture and, therefore, has ideological biases. Take the case of CollateX, a tool designed to compare texts with slight variations and align the parts of the texts that are different. Among other assumptions of the algorithm, it should be stressed that for CollateX it is not relevant to distinguish between a transposition or change of place of a portion of text (for example, in a poem, a stanza that appears displaced or in a different place) and a substitution (that is, elimination of a stanza from one place and the addition of the same lines in another place) (Van Zundert, in press). Here the question is not to establish whether CollateX's algorithm is correct. Researchers may or may not agree, but the key lies in knowing about this choice, this preference, and being aware that it conditions results and interpretations.

Digital models are thus icons that help us think and learn more about the original, the analogue object and the process of modelling is influenced by contextual elements.

Indeed, a few authors argue that theories and models are even more important in the era of Big Data because it is necessary to explain and understand the phenomena analysed through abstractions. In the Digital Humanities the concept of "model" is very widespread because it helps explain the core of digitisation work. Models are taken as tools, schemes or designs used in a specific context for particular purposes that are sometimes practical (to make a group of texts available online), but are often, especially in the academic field, speculative (to understand the structure of texts). More than the finished product, what matters in the Digital Humanities is the creative process that takes place when a phenomenon is "modelled", because the aim is to gain new knowledge, new meanings, by generating an external object that represents it.

The connection between the external object (for example, an epigraphic inscription) and the representation (a 3D reconstruction that allows the tombstone to be viewed from various angles and in greater detail) is based on similarity; it is therefore important to place reflection on "modelling" in context of the tradition of semiotics and the science of signs (Ciula and Eder, 2016). Naturally there are different degrees of similarity; the relationship can range from total likeness to metaphor, including a certain similarity between the properties of the object represented and the digital representation.



3D modelled epigraphic inscription. © Epigraphia 3D
http://www.epigraphia3d.es/

Digital models are thus icons that help us think and learn more about the original, the analogue object. This type of thought has been described as "abduction", because it stands somewhere between induction and deduction and is based on the intuition and experience of the person who "models" (Bryant and Raja, 2014). In other words, the process of modelling is influenced by contextual elements such as starting hypotheses, theoretical assumptions, scientific methods, formats and technologies.

Inside the Panopticon

The constant production of large amounts of data in real time through the social media also has a sinister counterpart. It is not unusual for Big Data to be compared to Big Brother or, better still, to the Panopticon – a type of penitentiary building devised by Jeremy Benthan in the eighteenth century which creates the sensation of being constantly watched – especially in the wake of the Edward Snowden case. Governments monitor citizens to ensure their security; this is by no means new and is part of the history of power structures studied by Michel Foucault, among others. In the modern state people are watched and, at the same time, encouraged to reveal their deepest secrets through confession, psychoanalytical therapy or, nowadays, by posting their "statuses" on Facebook.

As we have seen, the object of study of the humanities tends to be external, autonomous and finished – a historical document, a literary text, a visual representation – and research therefore does not usually pose ethical dilemmas on the privacy of creators and recipients. However, as consumers of culture, our acts are registered every time we search for a book, film or song on the internet, and when we click on a product and buy it; the same is true when we visit a museum – the surveillance camera is there to protect our heritage from crime and theft, but also to keep check of visitors; lastly, when we borrow a book from a public library a record is created in the database.

We should ask ourselves how humanists can study citizens' cultural habits, in constant dialogue with libraries and museums and using methods to anonymise data.

The case of public libraries is particularly interesting because they are a type of neighbourhood infrastructure accessible to everyone regardless of their economic status. Librarians record all loans, noting the date and borrower, in their databases. Nevertheless, this type of data is not accessible because municipal libraries have a long tradition of data protection (Starr, 2004). They do, however, publish lists of the most frequently borrowed books which function as indicators of contemporary taste. In order to be studied, this data would have to be published in an open format like XML or CSV and include a series of metadata such as the place and time of the loan, but such practices would encroach on users' privacy.

For researchers interested in reading habits, enjoying access to so much data would be a breakthrough. For example, it would be possible to ascertain how films, television and advertising influence people's tastes and reading habits. Manufacturers of electronic books, for example, are already using reading statistics to discover which books can be regarded as good – because readers finish them – despite not being best sellers; or to identify the next Dan Brown based on readers' degree of satisfaction with books written by unknown authors (Kobo, 2014). Basically, all the data generated by our electronic books is amassed by publishing companies to learn more about the relationship between sales and customer satisfaction; this makes it easier to justify economic decisions about the publishing future of a particular author, literary saga or genre.

By this I do not mean to imply that public libraries and museums should act in the same way as companies. I merely wish to point out that the state of being watched existed before the social media – just as spaces of resistance did. Just as companies like Twitter have been accused of exerting coercive power over research in the social sciences (Reichert, 2015), we should ask ourselves how humanists can study citizens' cultural habits, in constant dialogue with libraries and museums and using methods to anonymise data. In my view, we should aim to ensure that companies like Spotify and Amazon do not know more about a particular society – about our tastes, interests and moods – than its own members do.

Conclusions

Since 2011 the expression "Big Data" has been widely used in the experimental sciences and the media as if the increased amount of available data were the next scientific breakthrough. Although there is plenty of hype, the humanities have not been unaffected by this phenomenon; very specifically, although the digitisation of our cultural heritage is incomplete, several publications can be found which enter into conversation with Big Data and the social sciences. In European academia, there are many notable projects that process large amounts of data in order to study language, literature or art using techniques such as Natural Language Processing, automatic computer vision, topic modelling and stylometry.

After analysing the meaning of the expression "Big Data", this article highlights the cultural nature of data and defends the validity of theories, models and hypotheses for carrying out scientific research. Lastly, it discusses the dialectic between privacy and control. In a sense, this issue escapes the traditional field of the humanities, but it also deserves our attention as twenty-first-century citizens interested in the cultural practices of the present. Humanists no doubt have much to contribute to ethical and epistemological debates on the use of the data generated by citizens, recalling the "captured" and cultural nature of data, and bringing their experience to analysing particular cases bearing in mind the general context.

Bibliography

Anderson, Chris (06.23.08). "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". *Wired*. https://www.wired.com/2008/06/pb-theory/.

Berry, D. M. (2011). "The Computational Turn: Thinking about the Digital Humanities". *Culture Machine* 12. http://www.culturemachine.net/index.php/cm/article/viewarticle/440.

Blei, David M. (2012). "Probabilistic Topic Models". Communications of the ACM, 55.4: pp. 77–84: http://cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models/fulltext.

Bod, Rens (2013). *A New History of the Humanities*. Oxford University Press.

Boellstorff, Tom (2013). "Making Big Data, in Theory". *First Monday* 18.10. http://firstmonday.org/ojs/index.php/fm/article/view/4869.

Bryant, Anthony and Raja, Uzma (2014). "In the Realm of Big Data…" *First Monday*, 19.2. http://firstmonday.org/ojs/index.php/fm/article/view/4991.

Burrows, John (2002). "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship". *Literary and Linguistic Computing* 17.3: pp. 267–87.

Ciula, Arianna and Eder, Øyvind (2016). "Modelling in the Digital Humanities: Signs in Context". *Digital Scholarship in the Humanities*.

Drucker, Johanna (2011). "Humanities Approaches to Graphical Display". *DHQ: Digital Humanities Quarterly* 5.1. http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

English, James F. and Underwood, Ted (2016). "Shifting Scales: Between Literature and Social Science". *Modern Language Quarterly* 77.3: pp. 278–95. http://mlq.dukejournals.org/content/77/3/277.full.

Ganascia, Jean-Gabriel (2015). "Les *Big Data* dans les Humanités". *Critique* 818–19: pp. 627–36. https://www.cairn.info/revue-critique-2015-8-page-627.htm.

Ganascia, Jean-Gabriel, Glaudes, Pierre and Del Lungo, Andrea (2015). "Automatic Detection of Reuses and Citations in Literary Texts". *Literary and Linguistic Computing* 29.3: pp. 412–21.

Jänicke, S., Franzini, G., Faisal, C., Scheuermann, G. (2016). "Visual Text Analysis in Digital Humanities". *Computer Graphics Forum*, 35.2. DOI: 10.1111/cgf.12873

Jockers, Matthew (2013). *Macroanalysis. Digital Methods and Literary History*. University of Illinois Press.

Kobo, 2014. "Publishing in the Era of Big Data". http://news.kobo.com/_ir/159/20149/Publishing%20in%20the%20Era%20of%20Big%20Data%20-%20Kobo%20Whitepaper%20Fall%202014.pdf

Moretti, Franco (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso: London.

Nauta, Gerhard Jan and Heuvel, Wietske van den (2015). *Survey Report on Digitization in European Cultural Heritage Institutions 2015*. ENUMERATE. http://dataplatform.enumerate.eu/reports/survey-report-on-digitisation-in-european-cultural-heritage-institutions-2015/detail

Nudd, Tim (2016), *Spotify Crunches User Data in Fun Ways for This New Global Outdoor Ad Campaign*, *Adweek*. http://www.adweek.com/adfreak/spotify-crunches-user-data-fun-ways-new-global-outdoor-ad-campaign-174826

Reichert, Ramón (2015). "Big Data. Digital Media Culture in Transition". *Poetics and Politics of Data. The Ambivalence of Life in a Data-Driven Society*. Sabine Himmelsbach and Claudia Mareis (eds.). Basel, pp. 147–66.

Rosado Rodrigo, Pilar (2015). *Formas latentes: protocolos de visión artificial para la detección de analogías aplicados a la catalogación y creación artísticas*. Universitat de Barcelona. Doctoral thesis. http://www.tdx.cat/handle/10803/300302

Starr, Joan (2004). "Libraries and National Security: An Historical Review." *First Monday* 9.12. http://firstmonday.org/ojs/index.php/fm/article/view/1198/1118

Terras, Melissa (27. 03. 2013). "On Changing the Rules of Digital Humanities from the Inside". https://melissaterras.org/2013/05/27/on-changing-the-rules-of-digital-humanities-from-the-inside/

Van Zundert, Joris (in press). *Digital Scholarship in the Humanities*.

Ward, Jonathan Stuart and Barker, Adam (2013). "Undefined By Data: A Survey of Big Data Definitions". https://arxiv.org/abs/1309.5821.

Digital resources

1. Alliance of Digital Humanities Organizations: http://adho.org/
2. Europeana: http://www.europeana.eu/portal/es
3. European Association for Digital Humanities: http://eadh.org/
4. FreeLing: http://nlp.lsi.upc.edu/freeling/node/1
5. Asociación Humanidades Digitales Hispánicas: http://www.humanidadesdigitales.org/inicio.htm;jsessionid=FDC5ED5B005786714E45936B6E127DF8

6. Text Encoding Initiative: http://www.tei-c.org/index.xml
7. The Programming Historian: http://programminghistorian.org/
8. Stylo R: https://sites.google.com/site/computationalstylistics/stylo
9. Voyant: http://voyant-tools.org/
10. Google Arts & Culture: https://www.google.com/culturalinstitute/beta/u/0/?utm_campaign=cilex_v1&utm_source=cilab&utm_medium=artsexperiments&utm_content=freefall

Tweeters

1. Ted Underwood: @Ted_Underwood
2. Lev Manovich: @manovich
3. Nuria Rodríguez Ortega: @airun72
4. Greta Franzini: @GretaFranzini
5. Dev Verhoeven: @bestqualitycrab
6. Frank Fischer: @umblaetterer
7. Matthew Lincoln: @matthewdlincoln
8. José Calvo: @eumanismo
9. Elena González Blanco: @elenagbg
10. Dan Cohen: @dancohen